MPEG4 Audio

- MPEG-4 Audio Version 1 was finalized in October 1998 and published in 1999.
- MPEG-4 Audio Version 2 was finalized in December 1999 and published in 2000.
MPEG4 Audio Framework

- MPEG-4 standardizes natural audio coding at bitrates ranging from 2 kbit/s up to and above 64 kbit/s.
  - 2 - 4 kbit/s for speech with 8 kHz sampling
  - 4 - 16 kbit/s and for audio with 8 or 16 kHz sampling
  - Speech coding at medium bitrates 6 - 24 kbit/s uses Code Excited Linear Predictive (CELP). 8 and 16 kHz sampling rates are used.

The ITU-T speech coders currently operate at 6.3/5.3 (G.723), 8 kbps (G.729), 16 kbps (G.728), 32 kbps (G.721) 48/56/64 Kbps (G.722).

MPEG4 Audio

- Scalability
  - A given bitstream can be decoded by decoders of different levels of complexity. The audio quality, in general, is related to the complexity of the encoder and decoder used.

- Error Resilience
  - Error robustness provides the ability for a decoder to avoid or conceal audible distortion caused by transmission errors.
Audio Objects

- **Conversation object:**
The 'welcome' voice in this example is certainly the most important information. The speech is always located in front of the listener. This conversation may also be available in multiple languages.

- **Background object:**
The train will come from a distant location at the centre of the scene, pass the listener and then disappear behind him. In addition the low-frequency effect channel will reproduce a rumble noise. While the inclusion of this object is desirable, it can be removed in case of a very low bitrate connection.

- **Announcement object:**
For the announcement it is sufficient to transmit one low quality speech object. Some pseudo 3D and some echo effects can easily be generated at the MPEG-4 Player side.

- **Background Music:**
This orchestra may already be coded with MPEG-2 multichannel and the bitstream will be used without re-coding.

Different codecs for the different objects
Different bit rates for the different objects
Possibility to easily change the language of the speech
Different post processing on the different objects is possible
**MPEG4 audio Applications**

- Playing N-1 Audio Objects
- Multilingual Objects
- Movie Application
  - Conversation Object
  - Background Object
  - Announcement Object
  - Background Music

**MPEG4 Version 1**

- **Coding of Audio Objects (Audio Tools)**
  - Speech
    - Natural
      - HVXC (param.) (2 .. 4 kbit/s)
      - CELP (NB+WB) (4 .. 24 kbit/s)
    - Synthetic
      - TTS-Interface
  - General Audio
    - TwinVQ (6 .. 16 kbit/s/ch)
    - AAC (+scalable) (16 .. 64+ kbit/s/ch)
    - HILN

- **Composition of Audio Objects (System Tools)**
  - Audio scenes: Mixing, effects, …
MPEG4 Version 2

- Audio Tools:
  - Error Robustness
  - Low-Delay Audio Coding
    - Real time bi-directional communications
  - Small Step Scalability
    - Down to 1 Kbit/s from 16 Kbit/s in MPEG2-AAC
  - Parametric Audio Coding
  - CELP/HVXC Silence Compression
- New Systems Tools
  - Environmental Spatialisation

Error Robustness

- To be able to transmit on error-prone channels:
  - Unequal error protection
  - Error protection tools
    - Cyclic Redundancy Check
    - Forward Error Correction
  - Error resilience for source coding tools
MPEG4- AAC

- Extensions to MPEG2-AAC
  - Perceptual Noise Substitution (PNS)
  - Long Term Prediction
  - TwinVQ Coding Core
  - Bit Slice Arithmetic Coding (BSAC) (v2)
  - Low Delay AAC (v2)

Perceptual Noise Substitution (PNS)

- PNS permits a frequency selective parametric coding of noise-like signal components.
- Noise-like signal components are detected on a scalefactor band basis.
- Corresponding groups of spectral coefficients are excluded from quantization/coding. Instead, only a "noise substitution flag" plus total power of the substituted band is transmitted in the bitstream.
- Decoder inserts pseudo random vectors with desired target power as spectral coefficients
Tone-like signals require much higher coding precision than noise-like signals (e.g. 20 dB vs. 6 dB)

- Tonal signal components are predictable
- MPEG2 AAC:
  - Prediction of each spectral coefficient with backward adaptive predictor
  - High complexity (ca. 50% of decoder computation & RAM)
- MPEG4 AAC:
  - Long Term Predictor (LTP) as known from speech coding
  - Lower complexity: Saving of approx. 50% in terms of computation and memory over MPEG-2 predictors
  - Comparable performance to MPEG-2 predictors
Transform domain Weighted INterleave Vector Quantization (TwinVQ)

- Developed by NTT Japan
- TwinVQ doesn’t use any variable length coding and adaptive bit allocation, hence it has:
  - High coding gain for low bitrate (Typical 1:18 CR, but more CPU intensive than MP3)
  - Potential robustness against channel errors and packet loss
- It supports bitrate scalability, both by means of layered TwinVQ coding and in combination with the scalable AAC.

TwinVQ

- Vector selection under control of the perceptual model
**Bit Slice Arithmetic Coding**

- **Goal:** Smaller step scalability
  - For AAC the smallest enhancements layer are typically: 16 kbit/s.
- **BSAC** replaces AAC Huffman coding => 1 kbit/s/ch
- **Principle**
  - Transmit bit-slices with most significant bits first
  - Enhancement: less significant bits (finer quant.) higher frequency bands

**Low Delay AAC**

- **Sources of delay:**
  - **Frame length**
    - Reduced to 512 or 480 samples
    - Window shape has also been changed to lower overlap
  - **Filter bank delay**
    - Reduced just like the frame length
  - **Look-ahead time for block switching**
    - No block switching is performed
    - All pre-echo cancellation is handled by TNS
  - **Use of bit reservoir**
    - Is minimized in order to reach the desired delay (in an extreme case there is no bit reservoir)
Parametric Audio Coding

- For very low bitrate:
  - Signal decomposition into components:
    - "Harmonic and Individual Lines plus Noise" (HILN)
  - Functionalities:
    - very low bit rate (4 .. 16 kbit/s)
    - speed and pitch change (decoder)
      - Attractive for fast speech database search & browsing
    - bit rate scalability
      - 2.0kbps decoding is possible using 4.0kbps bit-stream

Parametric Audio Encoder
HVXC is a very efficient speech coding algorithm for low bit-rate speech:
- 1.2 kb/s to 1.7 kb/s var. rate
- 2.0 kb/s to 4.0 kb/s const. rate
- MPEG-4 Version 2 adds a variable Bitrate mode (VBR) mode to HVXC's capabilities.
- In MPEG4 HVXC and HILN coders are combined to provide very efficient parametric coding of speech and music at very low bit-rates, respectively.
Two different types of coding schemes are combined. One is suitable for voiced segments and the other for unvoiced segments.

- Voiced: Phase information is thrown away by harmonic representation of power spectrum of LPC residual.
  - Frequency domain analysis / synthesis.
- Unvoiced: Crisp consonant is obtained by CELP coding.
  - Time domain analysis / synthesis.
MPEG4 CELP

- CELP is one of the Speech coding algorithms in MPEG4:
  - Narrow band 3.85-12.2 kbps 10-40 ms frame
  - Wide band 10.9-23.8 kbps 10-20 ms frame
  - Multi-rate 200 - 800 bps step
  - Bit-rate scalability - 2.0kbps(NB), 4.0kbps(WB) step
  - Bandwidth scalability

Speech coders

<table>
<thead>
<tr>
<th>Signal Bandwidth, kHz</th>
<th>Channel Bitrate, kb/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.2</td>
<td>0</td>
</tr>
<tr>
<td>1.6</td>
<td>10</td>
</tr>
<tr>
<td>HYXC</td>
<td>20</td>
</tr>
<tr>
<td>CELP</td>
<td>21</td>
</tr>
<tr>
<td>WB CELP</td>
<td></td>
</tr>
</tbody>
</table>
Silence Compression

- In a bi-directional communication nearly 50% of the time there is silence.
  - We can generate noise at the decoder when there is no speech activity instead.

![Silence Compression Diagram]

Synthesized Sound

- Text to Speech (TTS)
  - Bit rate range from 200 bit/s to 1.2 Kbit/s
  - Speech synthesis using the prosody of the original speech
  - Lip synchronization control with phoneme information. (Face Animation)
  - Trick mode functionality: pause, resume, jump forward/backward.
  - International language and dialect support for text. (i.e., it can be signalled in the bitstream which language and dialect should be used)
  - International symbol support for phonemes, and support for specifying age, gender, speech rate of the speaker.
### TTS System

![Diagram of TTS System]

- **Text** → **Understanding**
- **Text-to-Phoneme Conversion** → **Prosody Generation**
- **Speech Generation** → **Synthesized Speech**

### TTS Applications

- **On-demand storytelling**
  - With or without Face Animation
  - Selectable age, gender, and speech rate
- **Motion picture dubbing with several levels of granularity**
  - Coarse: Aligning the composition time
  - Finely tuned: By sending information about the speaker lip
  - Finest: By using detailed prosody, and video related information
- **“Talking head” synthetic videoconferencing.**
Synthesized Sound

- Score Driven Synthesis
  - Using Structured Audio Orchestra Language (SAOL)
  - Control of the synthesis is accomplished by:
    - Scores: Sequenced set of commands that invokes various instruments at specific times
    - Scripts: Downloaded in a language called SASL (Structured Audio Score Language), can be used to create new sounds, and also include additional control information for modifying existing sound.
  - MPEG-4 does not standardize "a method" of synthesis, but rather a method of describing synthesis
    - Wavetable, FM, additive, physical-modeling, and granular synthesis, as well as non-parametric hybrids of these methods

MPEG4 Audio Summary

- MPEG-4 Audio provides tools for coding of both natural and synthetic audio objects.
- It permits to represent natural sounds (such as speech and music) and to synthesize sounds based on structured descriptions.
- The representation for synthesized sound can be derived from text data or so-called instrument descriptions and by coding parameters to provide effects, such as reverberation and spatialization.
- The representations provide compression and other functionalities, such as scalability or play-back at different speeds.
MPEG4 Encoding?

- Most of the techniques required for automatically producing a Structured Audio bitstream from an arbitrary sound are beyond today's state of the art.
- In the mean time, content authors will use special content creation tools to directly create Structured Audio bitstreams.
Dolby AC-3

- AC-3 (Audio Code number 3), also known as Dolby Digital, refers to a multichannel music compression technology that has been developed by Dolby Laboratories
  - Can carry from 1 to 5.1 channels.
    - Provides five full range channels (3 Hz to 20,000 Hz)
    - One bass-only effects channel (3 Hz to 120 Hz)
  - Input uncompressed PCM samples must be 32, 44.1 or 48 kHz on up to 20 bits.
  - Average Compression ratio 12:1

AC-3 Encoder

1. Overlapping blocks of 512 time samples are transformed into the frequency domain.
2. Due to the overlapping blocks, each PCM input sample is represented in two sequential transformed blocks. The frequency domain representation may then be decimated by a factor of two so that each block contains 256 frequency coefficients.
3. The individual frequency coefficients are represented in binary exponential notation as a binary exponent and a mantissa.
4. The set of exponents is encoded into a coarse representation of the signal spectrum which is referred to as the spectral envelope.
5. This spectral envelope is used by the core bit allocation routine which determines how many bits to use to encode each individual mantissa.
6. The spectral envelope and the coarsely quantized mantissas for 6 audio blocks (1536 audio samples) are formatted into an AC-3 frame.
AC-3 Encoder

References

- MPEG4 Audio FAQ: http://www.tnt.uni-hannover.de/project/mpeg/audio/faq/mpeg4.html
- MPEG4 Structured Audio: http://sound.media.mit.edu/~eds/mpeg4
- Synthetic and SNHC Audio in MPEG-4: http://leonardo.telecomitalialab.com/icjfiles/mpeg-4_si/10-SNHC_audio_paper/10-SNHC_audio_paper.htm
- MPEG Audio Resources and software: http://www.mpeg.org/MPEG/audio.html